Best Practices for Evaluating Method Validity, Data Quality and Study Reliability of Toxicity Studies for Chemical Hazard and Risk Assessments

Christopher Bevan[1] and Dale Strother[2]

[1]*CJB Consulting LLC, Loveland, OH 45140, USA*
[2]*ToxSolve LLC, Manassas, VA 20111, USA*

Risk assessments on chemicals are intended to provide sufficient information to evaluate the potential hazards associated with exposure to that chemical. It is important that risk assessments be based on the most reliable information available so that they can be used for sound decision-making by regulators and to convey appropriate risk information to the public. Although there are many uncertainties in risk assessments, these can be reduced by ensuring that the risk assessment reflects the most reliable and scientifically supported information on the potential hazards of the chemical, and that the quality and the completeness of the data are at a high standard. There are many different types of studies that a risk assessor needs to evaluate, and these vary not only in study design and methodologies, but also in quality. For the evaluation of the potential hazards to be credible, as well as objective and scientifically supported, a standardized procedure for data evaluation is desirable; this will also ensure consistency and transparency which are key factors in public acceptance. The risk assessor needs a process to ensure that data are of sufficient quality to underlie the hazard and risk assessments. This paper examines existing approaches or procedures used to evaluate data quality from *in vivo* and *in vitro* mammalian toxicity studies for use in hazard and risk assessments. Based on this review, the Klimisch approach with an expanded list of narrative descriptions should be used as a best practice for studies that have internationally accepted testing guidelines. The ToxRTool is proposed as the best practice to evaluate data quality for in vitro studies, as well as for studies that do not have internationally accepted guidelines.

1. Introduction and Guiding Principles

Standards for data quality have been most clearly established in programs for registration of chemical substances (*e.g.*, pharmaceuticals, food additives, pesticides, TSCA, REACH, *etc.*). In these cases, data development and quality are expected (or required by regulation) to conform to specific testing guidelines and Good Laboratory Practice (GLP) standards. Data quality, therefore, can be judged against these specific standards. Generally, a data quality evaluation procedure should consider several critical aspects such as method validity, reproducibility, and study reliability, in addition to the appropriateness or usefulness of the study for the assessment of human health risk. For example, the following specifics should be addressed in an evaluation of data quality: test substance identification; test system characterization; study design description; study results documentation; and biological plausibility of study design, data and results.

Advancing research and the development of new alternative testing approaches poses a constant challenge to the construct of standard guidelines for data development. This highlights the need for more attention to development of guiding principles for data quality evaluation.

In the event that key study data do not meet minimum requirements, additional studies may be needed to improve data quality. The presumption is that, for a given endpoint, data developed in accordance with accepted guidelines will carry more weight than data that do not. In the circumstance where data have been developed by methods not addressed by standard test guidelines, expert judgment is required to evaluate the quality and relevance of such data, with attention to the suitability for both hazard identification, and for quantification of risk. The expectation is that if an objective, transparent, well-documented justification for the use of such data is provided, such data will be accepted by independent peer reviewers.

In the case of studies which may not have prescriptive data development standards, the challenge of evaluating data quality is even greater. Data sets for existing chemicals may have been developed over a period of decades by a wide range of methods. Expert judgment and documentation of data quality evaluation become central in this case. Recent examples of programmatic efforts to review data for existing chemicals include the U.S. EPA and

OECD HPV programs and the EU REACH registration of existing chemicals. These programs have adopted the Klimisch scoring system for categorizing the reliability of test data (this is discussed in the next section).

2. Existing Practices and Guidance for Data Quality

The following section summarizes existing best practices and guidance for determining the data quality for hazard and risk assessments. Appendix A lists the sources searched and reviewed for information related to evaluation of data quality used in health risk assessment.

2.1. The Klimisch Evaluation System

2.1.1. Definitions

Klimisch *et al.* (1997) defined three aspects for evaluating data quality for hazard and risk assessments:

Reliability - evaluating the inherent quality of a test report or publication relating to preferably standardized methodology and the way the experimental procedure and results are described to give evidence of the clarity and plausibility of the findings.

Relevance - covering the extent to which data and/or tests are appropriate for a particular hazard identification or risk characterization.

Adequacy - defining the usefulness of data for risk assessment purposes. When there is more than one set of data for each effect, the greatest weight is attached to the study that is the most reliable and relevant.

2.1.2. Klimisch Codes of Reliability

Guidance for data evaluation based on the definitions given above has been developed for several hazard and risk assessment programs. These include: the OECD-SIDS program, the U.S. HPV program, and the EU REACH legislation, all using a scoring system developed by Klimisch *et al.* (1997) to assess the reliability of toxicology and ecotoxicology studies. The scoring system is also being used for evaluation of data on physico-chemical properties, and environmental fate and transport studies within these programs.

The Klimisch codes are defined as follows:

1 = reliable without restriction: "This includes studies or data from the literature or reports which were carried out or generated according to generally valid and/or internationally accepted testing guidelines (preferably performed according to GLP) or in which the test parameters documented are based on a specific (national) testing guideline (preferably performed according to GLP) or in which all parameters described are closely related/comparable to a guideline method."

2 = reliable with restrictions: "This includes studies or data from the literature, reports (mostly not performed according to GLP), in which the test parameters documented do not totally comply with the specific testing guideline, but are sufficient to accept the data or in which investigations are described which cannot be subsumed under a testing guideline, but which are nevertheless well documented and scientifically acceptable."

3 = not reliable: "This includes studies or data literature/reports in which there were interferences between the measuring system and the test substance or in which organisms/test systems were used which are not relevant in relation to the exposure (e.g. unphysiologic pathways of application) or which were carried out or generated according to a method which is not acceptable, the documentation of which is not sufficient for assessment and which is not convincing for an expert judgment."

4 = not assignable: "This includes studies or data from the literature, which do not give sufficient experimental details and which are only listed in short abstracts or secondary literature (books, reviews, *etc*.)."

Once a code has been assigned to a study, a short justification is provided, which documents the basis for the decisions for assigning a reliability code to the study data. In the paper by Klimisch *et al.* (2007), examples were given for justification statements used by the European Chemicals Bureau (ECB) in their software package for IUCLID version 2.12. These statements are listed in Table 1.

Table 1. Klimisch Criteria for Reliability Categories

Code	Justification
1	Guideline study (OECD, *etc*.)
	Comparable to guideline study
	Test procedure according to national standards (DIN, *etc*.)
2	Acceptable, well-documented publication/study report which meets scientific principles
	Basic data given; comparable to guidelines/standards
	Comparable to guideline study with acceptable restrictions
3	Method not validated
	Documentation insufficient for assessment
	Does not meet important criteria of today standard methods
	Relevant methodological deficiencies
	Unsuitable test system
4	Only short abstract available
	Only secondary literature (review, tables, books, *etc*.)

The European Centre for Ecotoxicology and Toxicology of Chemicals (ECETOC) has modified and expanded the justification phrases for each Klimisch reliability category for its Joint Assessment of Commodity Chemicals (JACC) program using alphanumeric scores (Table 2).

By assigning Klimisch codes to study data, the information for the hazard and risk assessment is ranked and organized for further review; this allows the risk assessor to focus on the most relevant studies, taking into account the endpoint being measured. Where there is more than one study for each endpoint, the greatest weight is attached to the studies that are the most relevant and reliable. The use of Klimisch codes institutes a data evaluation procedure based on method validity, study reliability and study quality using formal evaluation criteria with international acceptance.

The scoring of information, however, is not to exclude all unreliable data (Klimisch codes 3 and 4) from the hazard or risk assessment process. Some types of data that would not be acceptable because of some deficiency (*i.e.*, those where insufficient documentation exist for making an assessment) could be used as supporting data for either hazard identification or risk characterization. For instance, a weight-of-evidence approach could be used when there are several subchronic toxicity studies available on a particular chemical, none of which would be acceptable by itself due to some deficiency. Sound scientific judgment is needed in determining whether data from these supporting studies are adequate for a given requirement, along with adequate justification and documentation of the basis for the decision.

2.1.3. In vitro Studies

Internationally accepted guidelines exist for only a few *in vitro* study methods, primarily *in vitro* genotoxicity tests. The data from these *in vitro* genotoxicity tests are often used in hazard and risk assessment as alternatives to *in vivo* test data. But, in general, the results of *in vitro* tests are used as supporting or supplementary information in hazard and risk assessment for understanding the mode of action of a substance or in the interpretation of the relevance of animal data for humans.

Klimisch codes for reliability are of limited use for *in vitro* data because of the limited number of internationally accepted guidelines. Data quality evaluation of *in vitro* studies must consider both the adequacy and relevance of the data as well as the suitability of the methodology for a hazard or risk assessment.

Table 2. Criteria for Reliability Categories (modified by ECETOC)

Code	Category of reliability
1	Reliable without restriction
1a	'Good laboratory practice' guideline study (OECD, EC, EPA, FDA, *etc*.)
1b	Comparable to guideline study
1c	Test procedure in accordance with national standard methods (AFNOR, DIN, *etc*.)
1d	Test procedure in accordance with generally accepted scientific standards and described in sufficient detail
2	Reliable with restrictions
2a	Guideline study without detailed documentation
2b	Guideline study with acceptable restrictions
2c	Comparable to guideline study with acceptable restrictions
2d	Test procedure in accordance with national standard methods with acceptable restrictions
2e	Study well documented, meets generally accepted scientific principles, acceptable for assessment
2f	Accepted calculation method
2g	Data from handbook or collection of data
3	Not reliable
3a	Documentation insufficient for assessment
3b	Significant methodological deficiencies
3c	Unsuitable test system
4	Not assignable
4a	Abstract
4b	Secondary literature
4c	Original reference not yet available
4d	Original reference not translated
4e	Documentation insufficient for assessment

Criteria have been developed for what constitutes a fully validated *in vitro* assay (Appendix B). These criteria, which can be found in the OECD Guidance Document 34 (OECD, 2005), were initially established by the European Centre for the Validation of Alternative Methods (ECVAM) and the European Chemicals Bureau (ECB), and later refined by ECVAM (Hartung *et al.*, 2004). Klimisch *et al.* (2007) also listed criteria in their paper for data or information obtained from *in vitro* studies which are not carried out according to an international or national standard method. These criteria are:

- Description of the test system and test method in detail
- Purity/composition/origin of the test substance
- Data on the dose/concentration differentiated according to the toxicity of the test substance on the system; information on volatility
- Data on secondary effects which may influence a result (solubility, impurities, pH shifts, influence on the osmolarity, *etc.*)
- Appropriate negative/positive controls as integral parts of the test
- References on adequacy of the method should be given or generally known

2.*1.4. ECVAM ToxRTool Software*

A software-based tool called ToxRTool (Toxicological data Reliability Assessment Tool) has been developed by ECVAM to

provide criteria and guidance for reliability evaluations of both *in vivo* and *in vitro* toxicological data (Schneider *et al.*, 2009). The software is freely available to the public and a print-out of the Microsoft Excel® file is attached as Appendix C at the ECVAM website section "Publications" at http://ecvam.jrc.it.

The tool comprises of two parts, one for *in vivo* studies and the other for *in vitro* studies. There are five evaluation criteria groupings:

(1) test substance identification;

(2) test system characterization;

(3) study design description;

(4) study results documentation; and

(5) plausibility of study design and data.

For each criterion within a group, "1" is assigned if the study meets that criterion, or "0" if it is not met. A total of 21 points are possible for *in vivo* studies and 18 points for *in vitro* studies. The total number of points establishes a reliability category (the same as the Klimisch codes of reliability) based on the ranges presented in Table 3.

Building from the ToxRTool, Appendix D presents a draft list of elements (Becker, unpublished) that can help a risk assessor design a set of data evaluation procedures for determining the quality and reliability of an *in vitro* study. This list is not exhaustive and is not meant to be used as an inflexible checklist; certain elements will not be applicable for every type of in vitro study. Instead it is should be used only as a guide to help an assessor tailor the set of elements that will comprise the data evaluation criteria for addressing the specific set of studies under evaluation.

2.1.5 Limitations

Klimisch codes are used primarily for evaluating the reliability of study data. The question of the adequacy or usefulness of the data for the purpose of hazard and risk assessment can only be partially addressed by Klimisch codes. Klimisch codes would allow for weighing studies according to reliability, particularly where there are multiple studies where some or all of them have not been carried out according to current standard. However, the determination of the adequacy and relevance of information and which studies would be considered key or critical studies for the hazard and risk assessment require sound scientific judgment. The consequences of using information from studies conducted with exposure or experimental conditions not relevant to human exposure conditions can significantly influence the conclusions of a risk assessment.

2.2. U.S. Environmental Protection Agency (EPA) Guidance

2.2.1. U.S. EPA Data Quality Guidelines

In response to guidelines issued by the Office of Management and Budget (OMB, 2002), EPA published *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency* (USEPA, 2002), which describes EPA's policy and procedural guidance for ensuring and maximizing the quality of information that the Agency disseminates.

EPA's Science Policy Council (SPC) subsequently published five general assessment factors for evaluating the quality of scientific and technical information based on its existing information quality systems, practices and guidelines (USEPA, 2003). These assessment factors are described in Table 4 along with example questions provided by EPA.

Table 3. Reliability Category Output from ToxRTool Software (*See print-out in Appendix C)

Reliability Categorization (defined by Klimisch *et al.*, 1997)				**(Proposed) Consequence**
	In vivo	*In vitro*		
1	18-21	15-18	Reliable without restriction	Useful, check relevance for intended purpose
2	11-14	11-14	Reliable with restrictions	Potentially useful, check relevance for intended purpose
3	<11 or not all key criteria met*	<11 or not all key criteria met*	Not reliable	Generally not to be used as key study, but depending on the short-comings of the study, may still be useful in a weight-of evidence (WoE) approaches or as supportive information
4			Not assignable	Generally not to be used as key study, but depending on the short-comings of the study, may still be useful in WoE approaches or as supportive information (This category is not an outcome of this evaluation tool)

Table 4. EPA's Five General Assessment Factors for Evaluating Quality of Scientific Information

1. Soundness – The extent to which the scientific and technical procedures, measures, methods or models employed to generate the information are reasonable for, and consistent with, the intended purpose.

- Is the purpose of the study reasonable and consistent with its design?
- To what extent are the procedures, measures, methods or models employed to develop the information reasonable and consistent with sound scientific theory or accepted approaches?
- How do the study's design and results compare with existing scientific or economic theory and practice?
- Are the assumptions, governing equations and mathematical descriptions employed scientifically and technically justified?
- Is the study based on sound scientific or economic principles?

2. Applicability and Utility – The extent to which the information is relevant for the Agency's intended use.

- How useful or applicable is the scientific or economic theory applied in the study to the Agency's intended use of the analysis?
- How relevant are the study's purpose, design, outcome measures and results to the Agency's intended use of the analysis (e.g., for a chemical hazard characterization)?
- Are the domains (e.g., duration, species, exposure) where the model or results are valid useful to the Agency's application?
- How relevant is the study to current conditions of interest?

3. Clarity and Completeness – The degree of clarity and completeness with which the data, assumptions, methods, quality assurance, sponsoring organizations and analyses employed to generate the information are documented.

- To what extent does the documentation clearly and completely describe the underlying scientific or economic theory and the analytic methods used?
- To what extent have key assumptions, parameter values, measures, domains and limitations been described and characterized?
- To what extent are the results clearly and completely documented as a basis for comparing them to results from other similar tests?
- If novel or alternative theories or approaches are used, how clearly are they explained and the differences with accepted theories or approaches highlighted?
- Is the complete data set accessible, including metadata, data-dictionaries and embedded definitions (*e.g.*, codes for missing values, data quality flags and questionnaire responses)? Are there confidentiality issues that may limit accessibility to the complete data set?
- In the case of a modeling exercise, have the definitions and units of model parameters been provided?

Table 4. Continued - EPA's Five General Assessment Factors for Evaluating Quality of Scientific Information

4. Uncertainty and Variability - The extent to which the variability and uncertainty (quantitative and qualitative) in the information or the procedures, measures, methods or models are evaluated and characterized. • To what extent have appropriate statistical techniques been employed to evaluate variability and uncertainty? To what extent have the sensitive parameters of models been identified and characterized? • To what extent do the uncertainty and variability impact the conclusions that can be inferred from the data and the utility of the study? What are the potential sources and effects of error and bias in the study design? • Did the study identify potential uncertainties such as those due to inherent variability in environmental and exposure-related parameters or possible measurement errors?
5. Evaluation and Review - The extent of independent verification, validation and peer review of the information or the procedures, measures, methods or models. • To what extent has there been independent verification or validation of the study method and results? • What were the conclusions of these independent efforts, and are they consistent? • To what extent has independent peer review been conducted of the study method and results, and how were the conclusions of this review taken into account? • Has the procedure, method or model been used in similar, peer reviewed studies? Are the results consistent with other relevant studies? • In the case of model-based information, to what extent has independent evaluation and testing of the model code been performed and documented?

2.2.2. U.S. EPA FIFRA Pesticide Risk Assessments

Health and environmental risk assessments are conducted by the U.S. EPA as part of the pesticide registration process. As stated in the 40 CFR Part 158, U.S. EPA determines whether the data submitted or cited to fulfill the data requirements for the pesticide registration are acceptable. The determination is based on:

1. Design and conduct of the experiment from which the data were derived
- Whether generally accepted methods were used
- Sufficient numbers of measurements were made to achieve statistical reliability
- Sufficient controls were built into all phases of the experiment
- Whether the study was conducted in conformance with the design
- Whether good laboratory practices (GLP) were observed
- Reproducibility of the results

2. Evaluation of whether the data fulfill the purpose(s) of the data requirement.

EPA has published test guidelines that contain standards for conducting acceptable tests, guidance on the evaluation and reporting of data, definition of terms, and suggested study protocols. Any appropriate protocol may be used to generate the data required by this part, provided that it meets the purpose of the test standards specified in the pesticide assessment guidelines, and provides data of suitable quality and completeness as typified by the protocols cited in the guidelines. The toxicity studies submitted for a pesticide registration must be conducted according to GLP standards.

2.3. U.S. Food and Drug Administration (FDA) Guidance

2.3.1. FDA Food Additive Petition Review

Before a food additive may be used, the FDA must conclude that it is safe for its intended conditions of use. The petition must contain toxicity studies and other toxicology

information in order that a risk assessment can be carried out. The FDA has published its recommended protocols for all of the types of studies that may be required in a food additive petition, which is also known as the "Redbook 2000" (FDA, 2000). These toxicological studies must be carried out according to U.S. FDA GLP regulations.

The Redbook states that the toxicity test guidelines contained in the Redbook are only one possible approach in providing the toxicological basis for the safety assessment of the food additive. In these cases, FDA appears to be referring to modifications in the FDA-recommended protocols, and that these modifications should be proposed to FDA prior to conducting the study.

FDA's guidance for toxicity studies for food ingredients continues to emphasize the importance of sound scientific judgment.

2.4. European Food Safety Authority (EFSA) Guidance

The EFSA provides scientific advice for Europe's decision-makers in the areas of food and feed safety, nutrition, animal health and welfare, plant protection and plant health. In its risk assessments, the EFSA Scientific Panels decide what data sets are included or excluded in their assessment by determining these main aspects:

- Study design and power (*e.g.*, robust statistical design, potential bias)
- Data quality (*e.g.*, studies conducted in compliance with internationally agree guidelines)
- Relevance of the study for answering the specific question (*e.g.*, exposure assessment of (sub)populations, geographical regions, materials or test organisms used)
- Adequacy of data sets (*e.g.*, coverage of endpoints, sensitivity, specificity, appropriate statistical treatment of data, representativeness of data)
- Data sources (*e.g.*, peer-reviewed scientific literature, scientific reports, data bases, meeting abstracts)

Data of low quality is not to be disregarded completely, but the implications from those data should be considered, while taking into account the increased uncertainty caused by the reduced quality. The risk assessment is expected to explain the criteria for inclusion/exclusion of data, so that if data are excluded, this should be stated along with the rationale for their exclusion.

2.5 EU Environmental Risk Assessments for Pharmaceuticals Guidance

2.5.1. European Medicines Agency

In the EU an environmental risk assessment for veterinary and human pharmaceuticals is required as part of the safety submission for a Marketing Authorisation. The European Medicines Agency, the EU agency responsible for the scientific evaluation of these safety submissions, has provided some instructions for assessing reliability of data information in their guidelines on the environmental impact assessment for veterinary medicinal products (EMEA, 2008). In the Assessment Report, all of the studies that are summarized and evaluated are given a Reliability Index (RI) as a measure for the reliability; reliability is defined as the intrinsic reliability of a test with respect to the methodology and the description. These Reliability Indices are described in Table 5.

The EMEA (2008) states, however, that indicators of the "usefulness" of the data have not yet been developed.

Table 5. Reliability Indices Used in Environmental Impact Assessments for Veterinary Medicinal Products (EMEA, 2008)

Reliability Index (RI)	Definition	Description
1	Reliable	The methodology and the description are in accordance with the instructions in the EMEA guideline.
2	Less reliable	The methodology and/or the description are less in accordance with the instructions in the EMEA guideline.
3	Not reliable	The methodology and/or the description are not in accordance with the instructions in the EMEA guideline.

2.5.2 German Federal Environment Agency (UBA)

The German Federal Environment Agency (UBA) is responsible for the review of environmental risk assessment submissions of pharmaceuticals in Germany. The UBA has also developed criteria for assessing the data quality of published studies used in the environmental risk assessments of pharmaceuticals in Germany (Küster *et al.*, 2009). These criteria are a refinement of the data quality criteria in the EMEA (2008) guidance document and those criteria developed by Klimisch *et al.* (1997):

Category I:
Data are reliable without restriction according to the instructions in the EMA guideline and are therefore useable within the environmental risk assessment of pharmaceuticals. This category includes data from the literature or reports,

- Which were carried out or generated according to internationally accepted test guidelines (*e.g.* OECD, ISO).
- In which the test parameters documented are based on a specific (*e.g.* national) testing guideline (*e.g.* DIN).
- In which all parameters described are closely related/comparable to a guideline method.

Category II:
Data are reliable with restriction according to the instructions in the EMEA guideline and are useable within the environmental risk assessment of pharmaceuticals. This category includes data from the literature or reports,

- In which the test parameters documented do not totally comply with the respective test guideline, but are sufficient to evaluate the data.
- In which investigations are described which cannot be subsumed under a testing guideline, but which are nevertheless well documented, plausible and scientifically acceptable.

Category III:
Data are not reliable according to the instructions in the EMEA guideline and are not useable within the environmental risk assessment of pharmaceuticals. This category includes data from the literature or reports,

- Where the documentation is not sufficient for an assessment.
- Which were carried out or generated using a non-accepted method or which are not well documented, plausible and scientifically acceptable.
- Which are only listed in short information (*e.g.* abstracts, summaries, data safety sheets) or non-peer-reviewed literature (*e.g.* diploma/master thesis).

The highest grade of reliability is given to data from studies conducted and reported according to internationally accepted guidelines (*e.g.* OECD), and are used by UBA as reference standards. This also means that the standard methods recommended by the guidelines (*e.g.* OECD, ISO) are used as a reference. Literature studies might also be assessed as "reliable" if a literature report is available or if a test report is not performed according to standard methods, but is described sufficiently and carried out according to scientifically

acceptable standards. For these studies where a standard method was not used, detailed documentation should be provided and include data on the method and on the scope of the investigations, test conditions, measurements, and data analysis. Category III data are not used by UBA in the environmental risk assessments.

Küster *et al.* (2009) point out that in the UBA review studies with a higher reliability (*e.g.* valid test reports) have greater weight in the environmental risk assessment. Data with less reliability may be used as supplementary information especially if the results are comparable or in a similar range with results of higher reliability. Küster *et al.* (2009) also state "that within the review of pharmaceutical ERAs valid study reports outweigh reliable literature data from the public domain. Public literature, published scientific articles and reports that are publically available, may prevail, when a pharmaceutical, already on the market, has been investigated by scientists for whatever reason (*e.g.* as a model substance to investigate particular ecological processes)." These published studies are not necessarily conducted according to GLP and the raw data may not be available. UBA would consider these studies under the reliability category 2 in the environmental risk assessment review since these studies may provide valuable supplementary information.

3. Data Quality for Specific Endpoints

This section covers specific endpoints – genotoxicity and endocrine disruption - where there has been a focus not only on data quality and reliability, but also on the relevance of studies, and in particular the *in vitro* studies, for hazard and risk assessment.

3.1. Genotoxicity

The Risk Assessment Forum Technical Panel on Mutagenic Mode of Action within EPA provided a draft framework that could be used to determine whether available genotoxicity data on a certain chemical supported a mutagenic mode of action (MOA) for carcinogenicity (USEPA, 2007). It is an approach based on the Cancer Guidelines and Supplemental Guidance (USEPA, 2005a,b). The evaluation of data quality is an initial step in the MOA analysis, in which studies are judged against acceptability criteria and weighed for their appropriateness of use. Although EPA did not provide acceptability criteria, several publications were cited including the OECD test guidelines, the International Workshop on Genotoxicity Testing (IWGT, 1994; 2000; 2002; 2007), the International Conference on Harmonization (ICH, 1995; 1997), and other U.S. and international governmental agencies (Cimino, 2006).

Both *in vivo* and *in vitro* studies are used in hazard and risk assessments to derive conclusions about mutagen hazard and the mutagenic MOA for carcinogenicity. Chemicals that have positive findings in one or more *in vivo* studies supported by *in vitro* gene mutation or cytogenetic assays are generally considered to have a mutagenic MOA for carcinogenicity. Observation of effects *in vivo*, as contrasted with *in vitro*, tends to provide greater support for the conclusion, while the reverse may be less convincing.

It is recognized, however, that there may be cases where there are clear and reproducible positive responses seen in the *in vitro* genotoxicity studies, even when the *in vivo* studies are negative. The International Workshop on Genotoxicity Testing (IWGT) have worked on a framework for addressing appropriate interpretation and follow-up testing when initial *in vitro* tests are positive (Thybaud *et al.*, 2007). The IWGT emphasized the importance of considering the relevance to human health; for instance, whether other assays for the same endpoint support the results observed, whether the mode or mechanism of action is relevant to humans, and, most importantly, whether the effect observed is likely to occur *in vivo* at concentrations expected as a result of human exposure. The IWGT point out that *in vivo* assays are often mentioned by regulatory guidelines that may be conducted for follow-up testing. However, non-standard studies supported by peer-reviewed published

literature may, when justified, be more appropriate and informative than standard assays. The IWGT Working Group acknowledged that there are cases where mechanistic studies can be considered sufficient to support the indirect mode of action, and therefore additional *in vivo* data may not be necessary.

3.2. U.S. EPA Endocrine Disruption Screening Program (EDSP)

The Endocrine Disruption Screening Program (EDSP) was developed by the U.S. EPA to determine whether certain chemicals may have an effect in humans that is similar to an effect produced by a naturally occurring estrogen, or other such endocrine effects, using appropriate validated test systems and other scientifically relevant information. It is a two-tiered (Tier 1 and 2) screening and testing strategy, and a guidance document is available to provide basic principles and criteria for using a weight-of-evidence (WoE) approach for evaluation and interpretation of EDSP Tier 1 screening data to identify candidate chemicals for Tier 2 testing (USEPA, 2011). The Tier 1 assays, composed of both *in vitro* and *in vivo* screening assays, are specifically designed to determine the potential of a chemical to interact with the estrogen, androgen or thyroid hormonal pathways. The Tier 2 testing are *in vivo* studies that establish dose-response relationships for any potential adverse effects for the purposes of risk assessment.

In the evaluation of the Tier 1 data, the EPA guidance document (USEPA, 2011) describes a WoE analysis approach that includes evaluating the available data against current acceptance and quality criteria. The information is to be scientifically sound and relevant to addressing whether Tier 2 testing is needed.

EPA requires that EDSP Tier 1 assays need to be conducted using scientifically peer-reviewed study protocols. For each study, EPA will review the test methods employed and the conditions under which the studies (both guideline and non-guideline studies) were conducted to assess the standard of scientific quality, and thus, the level of confidence in the study findings for the WoE determination. The evaluation of individual EDSP Tier 1 assays or collection of assays/studies in the context of this WoE determination is facilitated by using the questions based on the considerations listed below to guide the analysis. These considerations are:

- The quality/validity of the method
- The reliability of the results
- The nature of the effect(s) observed
- Consistency and interrelationship among endpoints reported in an individual assay
- Relevance, specificity, and sensitivity of the endpoint(s) measured

In general, studies considered to be of good quality are those conducted according to scientifically acceptable methodology with sufficiently documented methods and data. Greater confidence will be given to data from studies that meet GLP standards, conducted using peer-reviewed test guidelines, or those studies meeting other quality assurance or standards.

Borgert *et al.* (2011) have published a detailed framework for conducting a hypothesis-based weight of evidence evaluation of EDSP Tier 1 results which is in many ways more transparent that the EPA EDSP WoE guidance. This approach requires constructing hypotheses for testing whether a substance interacts as an agonist or antagonist with components of estrogen, androgen, or thyroid pathways or with components of the aromatase or steroidogenic enzyme systems. In addition to recommending use of the Klimsich criteria, in the supplementary materials of this manuscript, the authors reference general scientific principles for evaluating data quality (discussed below in Section 5).

4. Proposed Best Practice(s)

Various procedures for evaluating data quality of *in vivo* and *in vitro* mammalian toxicity studies for hazard and risk assessments have been reviewed in this paper. A critical component for all of the reviewed evaluation procedures is the necessity for good documentation of studies (*i.e.*, test substance identification; test system characterization; study design description; study results; and biological plausibility of study design, data and results). This is accomplished by GLP standards for guideline toxicity studies, but GLP could equally apply to all experimental scientific studies. As pointed out by McCarty *et al.* (2012), GLP is not a unique component of guideline toxicity studies. It is a formal framework that ensures thorough, consistent, and detailed reporting of all aspects of experimental investigations so that reanalysis and reevaluation are readily possible.

There is no specific recommendation that will address data quality for all studies. For studies that have internationally accepted testing guidelines, the Klimisch system has been most widely accepted for evaluating data quality (reliability).

- ***It is proposed that the Klimisch approach be used with an expanded list of narrative descriptions such as, but not limited to, the ones used by ECETOC (see Table 2).***

The Klimisch system is not structured to evaluate studies that do not have internationally accepted guidelines. As discussed above, it is imperative that such studies be conducted according to scientifically acceptable methodology with sufficiently documented methods and data. A formal quality assurance process or standard, such as GLP, provides greater confidence in the data quality.

- ***It is proposed that an approach such as the ToxRTool be used to evaluate data quality for in vitro studies as well as for studies that do not have internationally accepted guidelines.***

5. Emerging Issues

Nontraditional toxicity tests and toxicity prediction tools are being considered for use in hazard and risk assessments of chemicals. Current methods for data quality evaluation are likely to be inadequate for ensuring a credible, as well as objective and scientifically supported, standardized procedure for these new techniques and approaches. As with traditional toxicity studies evaluation of the data will need a standardized procedure that will define acceptability criteria and how to weigh these studies for appropriateness of use in hazard and risk assessments. Considerable effort will be needed to define the critical parameters of these new tests and prediction tools to allow for data quality evaluation and to ensure confidence in the reliability and reproducibility of the data.

There is also increased recognition in the scientific community of the problems of data quality in the peer-review system of science publishing (Becker *et al.*, 2009; McCarty *et al.*, 2012; The Scientist, 2012). With the rapid pace of new scientific findings, the new technical advances that generate a tremendous amount of data, and the increased use of the Internet as a significant mode of gathering, communicating, and sharing of information, attention is being given to defining how emerging science by new and novel techniques should be evaluated and utilized in risk assessment.

Nevertheless, there are three fundamental tenets of data validity (data quality) by which all scientific data should be judged (Borgert *et al.*, 2011). First, the identity and authenticity of scientific measurements must be verifiable within a defined range of precision. Second, measurements and observations must not be confounded by extraneous factors and influences known to corrupt their accuracy and precision. Third, the measurements and observations must be replicable in independent hands.

Recognizing that new and novel methods, and studies with non-standard test systems can comprise an important part of body of the available scientific information for use in a

risk assessment, we suggest the following course of action for addressing such study reports. As Becker *et al.* (2009) point out, "Each study, GLP and non-GLP, should be evaluated and weighed in accordance with fundamental scientific principles. Factors to be evaluated include:
a) verification of measurement methods and data;
b) control of experimental variables that could affect measurements;
c) corroboration among studies;
d) power (both statistical and biological);
e) universality of the effects in validated test systems using relevant animal strains and appropriate routes of exposure;
f) biological plausibility of results; and
g) uniformity among substances with similar attributes and effects."

The first order task should be a thorough independent review of the methods and procedures and review of the data. The study documentation should be sufficient to permit independent verification/calculation of results, and the methods should be sufficiently described to permit replication of the study procedures by other laboratories. With the increasingly common practice of journals providing access to supplemental data, there are expanded opportunities for researchers to disseminate actual study data and this should facilitate independent evaluation by regulatory agencies. The determination of data quality and study reliability for such studies should reflect extent to which study details and actual data are available for independent verification.

6. Acknowledgements

Support for preparation of this white paper was provided by the Center for Advancing Risk Assessment Science and Policy of the American Chemistry Council. For more information on the Center see: http://arasp.americanchemistry.com/

References

Becker, R.A., Janus, E.R., White, R.D., Kruszewski, F.H., and Brackett, R.E. (2009) Good laboratory practices and safety assessments. Environ. Health Perspect. 117: A482-A483.

Borgert, C.J., Mihaich, E.M., Ortego,, L.S., Bentley, K.S., Homes, C.M., Levine, S.L., and Becker, R.A. (2011) Hypothesis-driven weight of evidence framework for evaluating data within the U.S. EPA's Endocrine Disruptor Screening Program. Regul. Toxicol. Pharmacol. 61: 185-191.

Cimino, M.C. (2006). Comparative overview of current international strategies and guidelines for genetic toxicology testing for regulatory purposes. Environ. Mol. Mutagen, 47: 362-390.

ECHA (2011) Guidance on Information Requirements and Chemical Safety Assessment. Chapter R.4: Evaluation of Available Information. European Chemicals Agency, Helsinki, Finland. Accessed online at http://echa.europa.eu/web/guest/guidance-documents/guidance-on-information-requirements-and-chemical-safety-assessment

EFSA (2009). Scientific Opinion: Transparency in Risk Assessment – Scientific Aspects. Guidance of the Scientific Committee on Transparency in the Scientific Aspects of Risk Assessment carried out by EFSA. Part 2: General Principles. The EFSA Journal 1051: 1-22.

EMEA (2008) Revised Guideline on Environmental Impact Assessment for Veterinary Medicinal Products in Support of the VICH Guidelines GL6 and GL38. European Medicines Agency. Committee for Medicinal Products for Veterinary Use (CVM), EMEA/CVMP/ERA/418282/2005-Rev.1, 17 Nov 2008. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/10/WC500004389.pdf.

FDA (2000). Toxicological Principles for the Safety Assessment of Food Ingredients (Redbook 2000), U.S. Food and Drug Adminstration; available at: http://www.fda.gov/Food/GuidanceComplianceRegulatoryInformation/GuidanceDocuments/FoodIngredientsandPackaging/Redbook/default.htm

FDA (2012) Guidance for Industry. S2(R1) Genotoxicity Testing and Data Interpretation for Pharmaceuticals Intended for Human Use, U.S. Food and Drug Administration. http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm074931.pdf

Hartung, T., Balls, M., Bardouille, C., Blanck, O., Coecke, S., Gstraunthaler, G., and Lewis, D. (2002) Good Cell Culture Practice. ECVAM Good Cell Culture Practice Task Force Report 1. ATLA 30: 407-414.

Hartung T., Bremer S., Casati S., Coecke S., Corvi R., Fortaner S., Gribaldo L., Halder M., Hoffmann S., Roi A., Prieto P., Sabbioni E., Scott L., Worth A. and Zuang V. (2004). A Modular approach to the ECVAM principles on test validity. Altern. Lab Anim. 32: 5: 467-472.

International Workshop on Genotoxicity Testing (IWGT). (1994) Reports from 1st IWGT. Mutat. Res. 312: 195-321.

International Workshop on Genotoxicity Testing (IWGT). (2000) Reports from 2nd IWGT. Environ. Mol. Mutagen. 35: 159-263.

International Workshop on Genotoxicity Testing (IWGT). (2002) Reports from 3rd IWGT. Mutat. Res. 540: 119-186.

International Workshop on Genotoxicity Testing (IWGT). (2007) Reports from 4th IWGT. Mutat. Res. 627: 1-118.

Klimisch, H.J., Andreae, M., and Tillmann, U. (1997). A systematic approach for evaluating the quality of experimental and toxicological and ecotoxicological data. Regulatory Toxicology and Pharmacology 25:1-5.

Küster, A., Bachmann, J., Brandt, U., Ebert, I., Hickmann, S., Klein-Goedicke, Maack, G., Schmitz, S., Thumm, E., and Rechenberg, B. (2009). Regulatory demands on data quality for the environmental risk assessment of pharmaceuticals. Regul. Toxicol. Pharmacol. 55: 276-280.

McCarty, L.S., Borgert, C.J., and Mihaich, E.M. (2012) Information quality in regulatory decision making: peer review versus good laboratory practice. Environ. Health Perspect. 120: 927-934.

OECD (2005). Guidance Document No. 34 on the Validation and International Acceptance of New or Updated Test Methods for Hazard Assessment. Organization for Economic Co-operation and Development.

OMB (2002). Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. (67 FR 8452). Accessed online at http://www.whitehouse.gov/omb/fedreg/reproducible2.pdf

Schneider, K., Schwarz, M., Burkholder, I., Kopp-Schneider, A., Edler, L., Kinsner- Ovaskainen, A., Hartung, T., and Hoffmann, S. (2009). ToxRTool, a new tool to assess the reliability of toxicological data. Toxicology Letters 189:138-144.

The Scientist (2012) Whither Science Publishing? The Scientist 26(08): 32-42.

Thybaud, V., AArdema, M., Clements, J., Dearfield, K., Galloway, S., Hayashi, M., Jacobson-Kram, D., Kirkland, D., MacGregor, J.T., Marzin, D., Ohyama, W., Schuler, M., Suzuki, H., and Zeiger, E. (2007). Strategy for genotoxicity testing: Hazard identification and risk assessment in relation to *in vitro* testing. Mutat. Res. 627: 41-58.

USEPA (2002). Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Environmental Protection Agency, EPA 260R-02-008, Office of Environmental Information. Available at http://www.epa.gov/quality/informationguidelines/documents/EPA_InfoQualityGuidelines .pdf

USEPA (2003). A summary of general assessment factors for evaluating the quality of scientific and technical information. Science Policy Council, Washington DC. EPA/100/B-03/001. Available at http://www.epa.gov/osa/spc/pdfs/assess2.pdf

USEPA (2005a) Guidelines for carcinogen risk assessment, EPA/630/P-03/001F, March 2005, (available at: http://www.epa.gov/cancerguidelines).

USEPA (2005b). Supplemental guidance for assessing cancer susceptibility from early-life exposure to carcinogens, EPA/630/R-03/003F, March 2005, (available at: http://www.epa.gov/cancerguidelines).

USEPA (2007). Framework for Determining a Mutagenic Mode of Action for Carcinogenicity: Using EPA's 2005 Cancer Guidelines and Supplemental Guidance for Assessing Susceptibility from Early-Life Exposure to Carcinogens, EPA 120/R-07/002-A, External Peer Review Draft, September 2007.

USEPA (2011). Endocrine Disruptor Screening Program. Weight-of-Evidence: Evaluating Results of EDSP Tier 1 Screening to Identify the Need for Tier 2 Testing. Office of Chemical Safety and Pollution Prevention, US Environmental Protection Agency, Washington DC, September 14, 2011. Available at http://www.epa.gov/endo/

Appendix A

The following sources and documents were searched and reviewed for information related to evaluation of data quality used in health risk assessment. Searches included documents containing the terms: “health risk assessment” and “best practices” or “data quality. Any procedures with a formal and documented process were included in the white paper.

- Google search engine
- PubMed database
- U.S. EPA Website
- OMB Website
- Australian Government Department of Health and Ageing website
  - Environmental Health Risk Assessment Guidelines for assessing human health risks from environmental hazards (2004)
- Canada Chemical Substance Risk Assessment Guidance Documents
- EPA Risk Assessment Guidance Documents
- European Chemicals Agency (ECHA) REACH Guidance Documents
- European Commission Scientific Committees website
- European Food Safety Authority website
- FDA Redbook 2000
- Guidance for Industry and Other Stakeholders: Toxicological Principles for the Safety Assessment of Food Ingredients
- National Academy of Sciences Reports
  - Science and Decisions: Advancing Risk Assessment (2009)
  - Toxicology Testing in the 21st Century (2007)
  - Science and Judgment in Risk Assessment (1994)
  - Issues in Risk Assessment (1993)
  - Risk Assessment in the Federal Government: Managing the Process (1983)
- OECD website
  - Manual for Investigation of HPV Chemicals
  - Guidance Documents
- WHO (2010) WHO Human Health Risk Assessment Toolkit: Chemical Hazards, IPCS Harmonization Project Document No.

Appendix B

The Criteria for Validation of *In vitro* Studies Derived from the OECD GD 34 (ECHA, 2008)

Concerned items	Decision criteria to be considered
Rationale for the test method	Clear statement of: - scientific basis - regulatory purpose - need for the test method
Relationship between the test method's endpoint and (biological) phenomenon of interest	Description of the scientific relevance of the measured effects Mechanistic (biological) or empirical (correlative) relationship to the specific type of effect or toxicity of interest
Detailed protocol for the test method	Detailed protocol and SOP including: - description of materials (including test substance) - what is measured - how it is measured - how data will be analyzed - decision criteria for evaluation - criteria for acceptable test performance
Test method performance using reference substances (accuracy assessment)	Sufficient number of reference substances measured in coded procedure Reference data and reference results for reference substances established
Performance evaluation	Performance evaluation in relation to: - relevant information from the species of concern - existing relevant toxicity testing data
Intra- and Inter-laboratory reproducibility	Data availability on - repeatability and reproducibility - robustness (variability)
Relevance	Demonstration of the predictive capacity of the method Precise definition of the applicability domain
Test method data quality	Evidence that all data supporting the validity are gained under quality conditions, *e.g.* GLP, GCCP
Data availability	All raw data should be available for expert review Detailed method protocol publically available

Appendix C

Print-out of the ECVAM ToxRTool Microsoft Excel® File

ToxRTool: a tool to assess the reliability of toxicological data		
Optional information for feed-back to ECVAM		
Date/period of evaluation:		
Rater Information		
Name		
Address		
Affiliation		
Education (subject and degree)		
Experience in toxicology (years)		
Experience in evaluating toxicological studies and/or performing risk assessments (years)		
Please answer these questions **after** having performed the evaluation!		
In general, do you consider a tool for assessing the reliability of toxicological data useful?		
Is the tool readily applicable and user-friendly?		
Is the tool transparent?		
How much time did you need (on average, or range) to read a case study and to apply the assessment tool to a study?	Reading/understanding (minutes per study	
	Tool application (minutes per study)	
Did the application of the assessment tool influence/change your opinion on how to assess data quality? If yes, in which way?		
Did you find it helpful to have the possibility to document observations on relevance of data (fields below reliability criteria)?		
Do you have specific comments on individual criteria (it is not necessary to repeat here the comments you might have given in the case study worksheets)?		
Do you have other/general comments on the tool?		

Print-out of the ECVAM ToxRTool Microsoft Excel® File (cont'd)

Explanations to ToxRTool

Objective: ToxRTool is designed to assess the inherent quality, also called reliability, of toxicological data as reported in a publication or a test report.

This tool essentially comprises a list of evaluation criteria. Criteria are subdivided in five groups:

I: Test substance identification, II: Test system characterization, III. Study design description,
IV: Study results documentation, V. Plausibility of study design and data

Per criterion either one ("1") or no ("0") point can be assigned. If a criterion is met, assign '1', if not assign '0'. Please choose from the respective drop-down list. **ALL CRITERIA MUST BE ANSWERED!**

In total 21 points for *in vivo* studies, 18 points for *in vitro* studies can be assigned. A reliability categorisation based on the total number of points is given below.

CRITERIA WRITTEN IN UPPER-CASE BOLD have special importance: points for each of the upper-case bold criteria are necessary to achieve Reliability category 1 or 2. **Please evaluate with special care!**

Data entry is requested in (and also restricted to) the shaded fields.

Reliability Categorisation (defined by Klimisch *et al.*, 1997)				**(Proposed) Consequence**
	In vivo	*In vitro*		
1	18-21	15-18	Reliable without restriction	Useful, check relevance for intended purpose
2	11-14	11-14	Reliable with restrictions	Potentially useful, check relevance for intended purpose
3	<11 or not all **UPPER-CASE BOLD** criteria met*	<11 or not all **UPPER-CASE BOLD** criteria met*	Not reliable	Generally not to be used as key study, but depending on the short-comings of the study, may still be useful in WoE approaches or as supportive information
4			Not assignable	Generally not to be used as key study, but depending on the short-comings of the study, may still be useful in WoE approaches or as supportive information (This category is not an outcome of this evaluation tool)

In addition to the criteria for assessing data reliability, at the bottom of the worksheets there are some questions to be answered optionally ("Optional documentation of observations with importance to relevance"). These questions allow to document observations in a non-formalised way, which may be of importance for the further use of the information for regulatory or other purposes.

Reliability assessment of *in vivo* toxicity studies

Study under evaluation

Authors:

Title:

Testing facility, year, sponsor, study no. or bibliographic reference:

Explanations are available for most criteria and show up, when the cursor is moved over the criteria field. Please read carefully! **UPPER-CASE BOLD criteria: the maximum score is needed for these criteria to achieve reliability category 1 or 2 (see Worksheet Explanations): Please evaluate with special care!**

Criteria			
No.	**Criteria Group I: Test Substance Identification**	**Score**	**Evaluator's explanations, comments on criteria, *etc.***
1	**WAS THE TEST SUBSTANCE IDENTIFIED?**		
2	Is the purity of the substance given?		
3	Is information on the source/origin of the substance given?		
4	Is all information on the nature and/or physico-chemical properties of the test item given, which you deem indispensable for judging the data (see explanations for examples)?		
		[Total]	
	Criteria Group II: Test organism characterisation		
5	**IS THE SPECIES GIVEN?**		
6	Is the sex of the test organism given?		
7	Is information given on the strain of test animals plus, if considered necessary to judge the study, other specifications (see explanation for examples)?		
8	Is age or body weight of the test organisms at the start of the study given?		
9	For repeated dose toxicity studies only (give point for other study types): Is information given on the housing or feeding conditions?		
		[Total]	

	Criteria Group III: Study design description		
10	**IS THE ADMINISTRATION ROUTE GIVEN?**		
11	**ARE DOSES ADMINISTERED OR CONCENTRATIONS IN APPLICATION MEDIA GIVEN?**		
12	**ARE FREQUENCY AND DURATION OF EXPOSURE AS WELL AS TIME-POINTS OF OBSERVATION EXPLAINED?**		
13	**WERE NEGATIVE (WHERE REQUIRED) AND POSITIVE CONTROLS (WHERE REQUIRED) INCLUDED (GIVE POINT ALSO, WHEN ASBSENT BUT NOT REQUIRED, SEE EXPLANATIONS FOR STUDY TYPES AND THEIR RESPECTIVE REQUIREMENTS ON CONTROLS)?**		
14	**IS THE NUMBER OF ANIMALS (IN CASE OF EXPERIMENTAL HUMAN STUDIES: NUMBER OF TEST PERSONS) PER GROUP GIVEN?**		
15	Are sufficient details of the administration scheme given to judge the study (see explanation for examples)?		
16	For inhalation studies and repeated dose toxicity studies only (given point for other study types): Were achieved concentrations analytically verified or was stability of the test substance otherwise ensured or made plausible?		
		[Total]	
	Criteria Group IV: Study results documentation		
17	Are the study endpoint(s) and their method(s) of determination clearly described?		
18	Is the description of the study results for all endpoints investigated transparent and complete?		
19	Are the statistical methods applied for data analysis given and applied in a transparent manner (give also point, if not necessary/applicable, see explanations)?		
		[Total]	
	Criteria Group V: Plausibility of study design and results		
20	**IS THE STUDY DESIGN CHOSEN APPROPRIATE FOR OBTAINING THE SUBSTANCE-SPECIFIC DATA AIMED AT (SEE EXPLANATIONS FOR DETAILS)?**		
21	Are the quantitative study results reliable (see explanations for arguments)?		
		[Total]	
	A Numerical result leads to initial Category:	[#]	
	B Checking UPPER-CASE BOLD scores leads to revised Category:	[#]	
	C Evaluator's proposal: Category:		
	D Justification in case evaluator deviates from B:		

	Optional documentation of observations with importance to relevance (not part of the reliability assessment)		
	During the course of the quality assessment observations may be made which are important for discussing the ***relevance*** *of the data for specific purposes. The optional possibility is provided here to document these observations for future use.*		
	What is the purpose of this quality evaluation (data documentation for use under REACH, classification activity under GHS, ECVAM validation activities, other)?		
	Study conducted according to recent OECD or EU guidelines (or other, *e.g.* national guidelines)? If yes, which ones? Study conducted under GLP conditions?		
	(If not a guideline study): Does a guideline exist for the study endpoint(s) in the study evaluated? If yes, which one?		
	Are you aware of relevant deviations from the guideline(s) in the study evaluated? If yes, which one?		
	Did you make observations with importance to the regulatory use of the data (example 1: evaluator may hint that a whole body inhalation study was performed with a substance, for which profound percutaneous absorption is expected or known, leading to substantial percutaneous uptake in addition to inhalation uptake; example 2: an Ames reversion assay was performed with strains able to identify frame-shift mutations only or without external metabolic activation; example 3: evaluator is in possession of positive evidence that the results obtained with the *in vitro* study under evaluation, in conjunction with known toxicokinetic data, are useful to assess the nephrotoxicity of the substance in humans)?		
	Would you like to make other/general comments on the usability of the data?		

Reliability assessment of *in vitro* toxicity studies

Study under evaluation

Authors:

Title:

Testing facility, year, sponsor, study no. or bibliographic reference:

Explanations are available for most criteria and show up, when the cursor is moved over the criteria field. Please read carefully! **UPPER-CASE BOLD criteria: the maximum score is needed for these criteria to achieve reliability category 1 or 2 (see Worksheet Explanations): Please evaluate with special care!**

Criteria			
No.	**Criteria Group I: Test Substance Identification**	**Score**	**Evaluator's explanations, comments on criteria, *etc.***
1	**WAS THE TEST SUBSTANCE IDENTIFIED?**		
2	Is the purity of the substance given?		
3	Is information on the source/origin of the substance given?		
4	Is all information on the nature and/or physico-chemical properties of the test item given, which you deem indispensable for judging the data (see explanations for examples)?		
		[Total]	
	Criteria Group II: Test organism characterisation		
5	Is the test system described?		
6	Is information given on the source/origin of the test system?		
7	Are necessary information on test system properties, and on conditions of cultivation and maintenance given?		
		[Total]	
	Criteria Group III: Study design description		

8	Is the method of administration given (see explanation for details)?		
9	**ARE DOSES ADMINISTERED OR CONCENTRATIONS IN APPLICATION MEDIA GIVEN?**		
10	**ARE FREQUENCY AND DURATION OF EXPOSURE AS WELL AS TIME-POINTS OF OBSERVATION EXPLAINED?**		
11	**WERE NEGATIVE CONTROLS INCLUDED (GIVE ALSO POINT, IF NOT NECESSARY, SEE EXPLANATIONS)?**		
12	**WERE POSITIVE CONTROLS INCLUDED (GIVE ALSO POINT, IF NOT NECESSARY, SEE EXPLANATIONS)?**		
13	Is the number of replicates (or complete repetitions of experiment) given?		
		[Total]	
	Criteria Group IV: Study results documentation		
14	Are the study endpoint(s) and their method(s) of determination clearly described?		
15	Is the description of the study results for all endpoints investigated transparent and complete?		
16	Are the statistical methods for data analysis given and applied in a transparent manner (give also point, if not necessary/applicable, see explanations)?		
		[Total]	
	Criteria Group V: Plausibility of study design and results		
17	**IS THE STUDY DESIGN CHOSEN APPROPRIATE FOR OBTAINING THE SUBSTANCE-SPECIFIC DATA AIMED AT (SEE EXPLANATIONS FOR DETAILS)?**		
18	Are the quantitative study results reliable (see explanations for arguments)?		
		[Total	
	A Numerical result leads to initial Category:	[#]	
	B Checking UPPER-CASE BOLD scores leads to revised Category:	[#]	
	C Evaluator's proposal: Category:		
	D Justification in case evaluator deviates from B:		
	Optional documentation of observations with importance to relevance (not part of the reliability assessment)		
	During the course of the quality assessment observations may be made which are important for discussing the ***relevance*** *of the data for specific purposes. The optional possibility is provided here to document these observations for future use.*		

	What is the purpose of this quality evaluation (data documentation for use under REACH, classification activity under GHS, ECVAM validation activities, other)?		
	Study conducted according to recent OECD or EU guidelines (or other, *e.g.* national guidelines)? If yes, which ones? Study conducted under GLP conditions?		
	(If not a guideline study): Does a guideline exist for the study endpoint(s) in the study evaluated?		
	Are you aware of relevant deviations from the guideline(s) in the study evaluated? If yes, which one?		
	Did you make observations with importance to the regulatory use of the data (example 1: evaluator may hint that a whole body inhalation study was performed with a substance, for which profound percutaneous absorption is expected or known, leading to substantial percutaneous uptake in addition to inhalation uptake; example 2: an Ames reversion assay was performed with strains able to identify frame-shift mutations only or without external metabolic activation; example 3: evaluator is in possession of positive evidence that the results obtained with the *in vitro* study under evaluation, in conjunction with known toxicokinetic data, are useful to assess the nephrotoxicity of the substance in humans)?		
	Would you like to make other/general comments on the usability of the data?		

Appendix D

List of Elements to be Reviewed for Designing a Set of Data Evaluation Procedures for Determining the Quality and Reliability of *In vitro* Studies

Test Article Identity and Purity
• Was the source of the test substance identified?
• Was the purity of the substance specified?
• Was the purity of the substance independently verified at start of the test?
• Were the concentrations of test substance in dose matrix verified by analysis?
• Source of vehicle identified? Purity of vehicle specified, etc.
• Are the physical chemical properties of the test material and solvents compatible with the culture media/incubation conditions/test conditions?
• Was the possibility of the test material reacting with culture vessel (sticking to walls) and altering culture conditions (e.g., changes in pH and osmolarity) considered/controlled for?
• Are all of the Culture media components described?
• Are the test material compatible with the assays (e.g., inhibit enzyme activities etc.)?
Test organism
• Was test species/strain/sex/age identified/cell type/ Lineage/origin/passage number / source identified?[a]
• For primary cells, was the source and host (demographics) identified?[a]
• For stem cells, was validated methodology used?
• Were there records of equipment calibration (incubator, water baths, pipets etc.)?
• Were appropriate acclimation procedures used and described?
• Were instruments calibrated appropriately prior to use?
• Were the treatment methods adequately described (dose concentrations, solvents and methods of culture addition)
• Were observations of cells during the study for signs of viability, infection or other observations appropriately conducted?
• Was cytotoxicity controlled for?
• Was the cell cycle phase identified and controlled for?
• Was the karyotype identified and controlled for?
• Was the cell density identified and controlled for?
• Was the cell phenotype (*e.g.*, state of differentiation; expression of specific activities; correct receptor expressed) verified?
Study Design and Conduct: Was the study design appropriate for the questions evaluated?
• Was there a complete written study protocol unique for the specific investigation (description of all of the relevant elements of the study)

• Tests system (guideline or non-guideline study and species, strain, sex, age, appropriate route of exposure) *etc.*
• Were standardized and well validated test methods for endpoints examined used?
• Appropriate Dose levels selected • Full dose response assessment
• Were appropriate calibration, standard curves, etc. conducted
• Appropriate numbers / replicates in each treatment group
• Appropriate instruments and procedures for data collection used
• Appropriate negative concurrent controls
• Appropriate concurrent positive controls (always necessary)
• Data collection methods (are raw data or underlying data available?)
• Are test result materials (images, raw data records, etc.) available for independent audit
• Were deviations from the study protocol documented and made part of the study records and reports?
• Are records available documenting technician training in the procedures employed?
• Are records available documenting that all aspects of the study were conducted in accordance with a quality control plan and that the records and reports have been quality assured?
Data Analysis methods
• Were appropriate statistical analyses used?
• Were all data included in the stats analyses (or exclusion statistically supported/justified scientifically)?
• Are underlying data available to independently review and verify the reported summary statistics
• Are underlying data available to conduct alternative statistical evaluations and comparison of the results?
• Sample replicates/numbers per group
• Statistical power

[a]Hartung *et al.* (2002)